A Simple yet Effective Method for Sentence Ordering

Aili Shen **Timothy Baldwin**
The University of Melbourne
{aili.shen, tbaldwin}@unimelb.edu.au

Abstract

Sentence ordering is the task of arranging a given bag of sentences so as to maximise the coherence of the overall text. In this work, we propose a simple yet effective training method that improves the capacity of models to capture overall text coherence based on training over pairs of sentences/segments. Experimental results show the superiority of our proposed method in in- and cross-domain settings. The utility of our method is also verified over a multi-document summarisation task.

1 Introduction and Background

Document coherence understanding plays an important role in natural language understanding, where a coherent document is connected by rhetorical relations, such as *contrast*, *elaboration*, *narration*, and *justification*, allowing us to communicate cooperatively in understanding one another. In this work, we measure the ability of models to capture document coherence in the strictest setting: sentence ordering (Barzilay and Lapata, 2005; Elsner et al., 2007; Barzilay and Lapata, 2008; Prabhumoye et al., 2020), a task of ordering an unordered bag of sentences from a document, aiming to maximise document coherence.

The task of sentence ordering is to restore the original order for a given bag of sentences, based on the coherence of the resulting document. The ability of a model to reconstruct the original sentence order is a demonstration of its capacity to capture document coherence. Figure 1 presents such an example, where the (shuffled) sentences are from a paper abstract discussing the relationship between word informativeness and pitch prominence, and the gold-standard sentence ordering is (4, 5, 1, 7, 3, 2, 6). Furthermore, the task of sentence ordering is potentially beneficial for downstream tasks such as multi-document summarisation (Nallapati et al., 2017), storytelling (Fan et al., 2019; Hu et al., 2020), cooking recipe generation (Chandu et al., 2019), and essay scoring (Tay et al., 2018; Li et al., 2018), where document coherence plays an important role.

(1) But there are others who express doubts about such a correlation.
(2) They also show that informativeness enables statistically significant improvements in pitch accent prediction.
(3) Our experiments how that there is a positive correlation between the informativeness of a word and its pitch accent assignment.
(4) In intonational phonology and speech synthesis research, it has been suggested that the relative informativeness of a word can be used to predict pitch prominence.
(5) The more information conveyed by a word, the more likely it will be accented.
(6) The computation of word informativeness is inexpensive and can be incorporated into speech synthesis systems easily.
(7) In this paper, we provide some empirical evidence to support he existence of such a correlation by employing two widely accepted measures of informativeness.

Figure 1: An example of shuffled sentences from the same document.

Traditional approaches to sentence ordering used hand-engineered features to capture document coherence (Barzilay and Lapata, 2005; Elsner et al., 2007; Barzilay and Lapata, 2008; Elsner and Charniak, 2011; Mesgar and Strube, 2016), e.g. using an entity matrix (Barzilay and Lapata, 2005, 2008) or graph (Guinaudeau and Strube, 2013) to represent entity transitions across sentences, and maximising transition probabilities between adjacent sentences.

Neural work has modelled the task either generatively (Li and Hovy, 2014; Li and Jurafsky, 2017; Gong et al., 2016; Logeswaran et al., 2018; Cui et al., 2018; Wang and Wan, 2019; Oh et al., 2019; Cui et al., 2020; Yin et al., 2020; Kumar et al., 2020) or discriminatively (Chen et al., 2016; Prabhumoye et al., 2020). As example genera-

Proceedings of the 22nd Annual Meeting of the Special Interest Group on Discourse and Dialogue, pages 154–160
July 29–31, 2021. ©2021 Association for Computational Linguistics

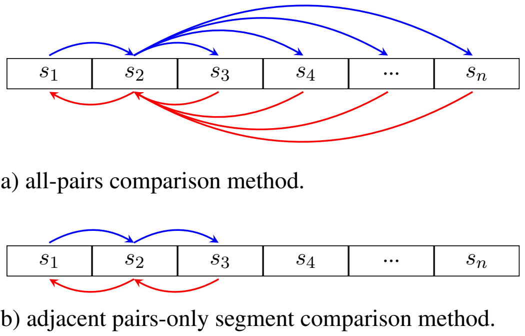
tive approaches, Cui et al. (2020) obtain sentence and paragraph representations from BERT (Devlin et al., 2019) and then use a pointer network to decode the sentence ordering for a given paragraph, whereas Yin et al. (2019) use a graph-based neural network over sentences and entities. The shortcoming of generative methods is the difficulty in obtaining good paragraph representations, especially for longer paragraphs. To mitigate this, various attention mechanisms have been explored (Cui et al., 2018; Wang and Wan, 2019; Kumar et al., 2020).

Discriminative approaches, on the other hand, can readily capture the relative order between sentence pairs, and paragraph decoding can then be achieved through methods such as beam-search (Chen et al., 2016) or topological sort (Tarjan, 1976; Prabhumoye et al., 2020). However, even with exact decoding methods such as topological sort, issues remain, including: (1) coherence scores for sentence pairs that are distant in the document tend to be noisy; and (2) it can be difficult to determine the relative order of adjacent sentences without broader context. To mitigate these two drawbacks, we propose a simple yet effective training method. Instance pairs are only constructed from adjacent segments to provide stronger coherence signals, but to capture broader context, up to 3 continuous sentences are combined to form a single segment in an instance pair. The effectiveness of our method is demonstrated across multiple datasets, in in- and cross-domain settings, and the setting of multi-document summarisation.

2 Methodology

The method proposed by Prabhumoye et al. (2020) exploits the relative order between *any two* sentences in a given paragraph. As in Figure 2a, the pairs connected by blue and red lines (pointing right and left, resp.) are the resulting positive and negative coherence instances for sentence s_2, respectively. These instances are used to train a text coherence model, which we denote as “allpairs”.

In contrast, our method utilises the relative order between *adjacent* segments only, resulting in an order of magnitude less training data than allpairs ($\mathcal{O}(n)$ vs. $\mathcal{O}(n^2)$) but stronger supervision signal; we denote this as “adjonly”. As in Figure 2b, the blue/red lines connect adjacent sentences for sentence s_2, resulting in positive/negative coherence instances. To capture broader context, we also construct pairs based on segments made up of multiple continuous sentences (not shown in the figure), such as ($s_{1:2}$, $s_{2:3}$) and ($s_{1:3}$, $s_{2:4}$) as positive instances, and ($s_{2:3}$, $s_{1:2}$) and ($s_{2:4}$, $s_{1:3}$) as negative instances, where $s_{i:i+j}$ denotes the concatenation of sentences s_i to s_{i+j} inclusive ($j \geq 0$). In this work, we experiment with $j \in \{0, 1, 2\}$ (i.e. sentence unigrams, bigrams, and trigrams), resulting in (at most) $6(n-2)$ instances for a paragraph with n sentences (noting that the segment cannot extend beyond the extremities of the document).



(a) all-pairs comparison method.

(b) adjacent pairs-only segment comparison method.

Figure 2: Illustration of the baseline method of Prabhumoye et al. (2020) (a) and our proposed training method (b), where blue and red lines indicate positive and negative segment pairs, respectively.

At test time, following Prabhumoye et al. (2020), we predict the relative order of each sentence pair (only sentence unigram), then order the sentences with topological sort.

We also trialled other training methods — including regressing over the distance between two sentences, and training with constraints over sentence triplets inspired from Xu et al. (2019a) in computer vision — but observed no improvement.

3 Experiments

3.1 Datasets

We perform experiments over six publicly available datasets from Logeswaran et al. (2018) and Xu et al. (2019b), resp.:

- **NeurIPS, ACL, and NSF:** abstracts from NeurIPS papers, ACL papers, and NSF grants (ave. sentences = 6.2, 5.0, and 8.9, resp.).
- **Athlete, Artist, and Institution:** paragraphs with >10 sentences from Wikipedia articles of athletes, artists, and educational institutions (ave. sentences ≈ 12).

3.2 Evaluation Metrics

Following previous work, we use 4 evaluation metrics (higher is better in each case):

- **Perfect Match Ratio (PMR):** % of paragraphs for which the entire sequence is correct (Chen et al., 2016).
- **Accuracy (Acc):** % of sentences whose absolute positions are correct (Logeswaran et al., 2018).
- **Longest Common Subsequence (LCS):** % overlap in the longest common subsequence between the predicted and correct orders (Gong et al., 2016).
- **Kendall's Tau (τ):** rank-based correlation between between the predicted and correct order (Lapata, 2006).

3.3 Model Configuration

We benchmark against Prabhumoye et al. (2020), using a range of text encoders, each of which is trained separately over allpairs and adjonly data.

LSTM: each segment is fed into a separate biLSTM (Hochreiter and Schmidhuber, 1997) with the same architecture and shared word embeddings to obtain representations, and the segment representations are concatenated together to feed into a linear layer and softmax layer. We use 300d pre-trained GloVe word embeddings (Pennington et al., 2014) with updating, LSTM cell size of 128, and train with a mini-batch size of 128 for 10 epochs (with early stopping) and learning rate of 1e-3.

BERT: predict the relative order from the "CLS" token using pre-trained BERT (Devlin et al., 2019), or alternatively ALBERT (Lan et al., 2020) (due to its specific focus on document coherence) or SciBERT (Beltagy et al., 2019) (due to the domain fit with the datasets). For BERT and ALBERT, we use the base uncased version,[1] and finetune for 2 epochs in each case with a learning rate of {5e-5, 5e-6}.

BERTSON (Cui et al., 2020): the current SOTA for sentence ordering, in the form of a BERT-based generative model which feeds representations of each sentence (given the context of the full document) into a self-attention based paragraph encoder to obtain the document representation, which is used to initialise the initial state of an LSTM-based pointer network. During decoding, a deep relational module is integrated with the pointer network, to predict the relative order of a pair of sentences.[2]

3.4 In-domain Results

Table 1 presents the results over the academic abstract datasets. The adjacency-only method performs better than the all-pairs method for all encoders over all evaluation metrics, underlining the effectiveness of our proposed training method. Comparing sentence encoders, the pretrained language models outperform LSTM, with ALBERT and SciBERT generally outperforming BERT by a small margin, demonstrating the importance of explicit document coherence training (ALBERT) and domain knowledge (SciBERT). Overall, SciBERT-adjonly achieves the best over NeurIPS and ACL, and ALBERT-adjonly achieves the best over NSF.

As BERTSON is trained on BERT base, the fairest comparison is with BERT-adjonly. Over NeurIPS, BERTSON has a clear advantage, whereas the two models are perform almost identically over ACL, and BERT-adjonly has a clear advantage over NSF. Note that this correlates with an increase in average sentence length (NSF $>$ ACL $>$ NeurIPS), suggesting that our method is better over longer documents.

Looking to the results over the Wikipedia datasets in Table 2, once again the adjacency-only model is consistently better than the all-pairs method. Here, ALBERT-adjonly is the best of BERT-based models (noting SciBERT has no domain advantage in this case), and despite the documents being longer again than NSF on average, there is remarkable consistency with the results in Table 1 in terms of the evaluation metrics which are explicitly normalised for document length (LCS and τ).

3.5 Cross-domain Results

To examine the robustness of our method in a cross-domain setting, we focus exclusively on ALBERT, given its overall superiority in an in-domain setting. We finetune ALBERT over the Athlete dataset, and test over the Artist, Institution, and NeurIPS datasets, resulting in different degrees of topic and domain shift: Athlete → Artist (similar

[1] For SciBERT, we use scivocab base uncased version, where the vocabulary is based on scientific text.

[2] Note that the code for BERTSON has not been released, and given the complexity of the model, we were not confident of our ability to faithfully reproduce the model. As such, we only report on results from the paper, for those datasets it was evaluated over. Similar to Prabhumoye et al. (2020), all sentence pairs are used to learn the sentence representations, aiming to capture the pairwise relationship between sentences.

Models	NeurIPS				ACL				NSF			
	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ
BERTSON	**48.01**	**73.87**	—	0.85	59.79	78.03	—	0.85	23.07	50.02	—	0.67
LSTM-allpairs	14.18	43.62	71.58	0.66	26.76	50.19	75.05	0.66	6.05	23.20	56.82	0.48
LSTM-adjonly	18.16	47.10	74.44	0.69	30.66	53.08	76.94	0.70	9.34	34.98	67.36	0.65
BERT-allpairs	33.83	61.91	83.10	0.82	50.34	69.35	85.94	0.83	14.43	38.58	71.05	0.70
BERT-adjonly	42.29	68.06	86.23	0.85	59.79	75.96	89.72	0.86	23.24	54.23	81.12	0.81
ALBERT-allpairs	37.31	65.12	85.00	0.83	54.01	71.71	87.36	0.85	14.33	38.79	71.22	0.70
ALBERT-adjonly	41.79	68.95	86.23	0.84	60.97	76.40	90.09	0.87	**25.34**	**56.71**	**82.62**	**0.82**
SciBERT-allpairs	37.31	65.55	84.65	0.84	54.74	72.23	87.40	0.85	14.84	39.56	71.80	0.71
SciBERT-adjonly	44.53	71.00	**87.74**	**0.87**	**63.04**	**78.98**	**90.87**	**0.89**	24.65	55.91	82.18	**0.82**

Table 1: Results over the academic abstract datasets (results for BERTSON are those reported in Cui et al. (2020); "—" indicates the number was not reported in the original paper).

Models	Athlete				Artist				Institution			
	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ
LSTM-allpairs	0.00	15.31	49.32	0.28	0.00	12.62	46.23	0.20	9.04	28.59	58.47	0.40
LSTM-adjonly	0.89	30.54	64.91	0.63	0.00	24.32	60.24	0.51	21.16	45.56	72.07	0.70
BERT-allpairs	2.53	32.81	68.24	0.63	0.66	24.45	61.16	0.50	22.01	43.94	71.85	0.64
BERT-adjonly	10.17	50.52	79.56	0.79	6.93	46.59	76.82	0.76	25.94	56.12	80.60	0.79
ALBERT-allpairs	2.78	35.03	69.99	0.65	1.23	29.57	66.25	0.59	21.84	47.64	75.19	0.71
ALBERT-adjonly	**14.89**	**56.25**	**82.59**	**0.82**	**9.31**	**49.66**	**79.64**	**0.78**	**28.50**	**58.86**	**82.93**	**0.81**
SciBERT-allpairs	1.14	27.97	64.47	0.56	0.38	22.36	59.72	0.47	17.41	40.06	70.11	0.61
SciBERT-adjonly	6.08	45.40	76.27	0.75	2.18	39.42	72.40	0.71	21.33	51.71	77.96	0.77

Table 2: Results over the Wikipedia datasets.

Models	Artist				Institution				NeurIPS			
	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ	PMR	Acc	LCS	τ
ALBERT-allpairs	1.14	29.37	66.15	0.58	0.34	26.69	64.12	0.54	20.90	49.57	76.18	0.66
ALBERT-adjonly	**8.83**	**48.74**	**78.93**	**0.78**	**4.78**	**41.43**	**74.31**	**0.72**	**35.82**	**61.41**	**83.29**	**0.78**

Table 3: Cross-domain results, with finetuning over the Athlete dataset.

topic), Athlete → Institution (topic change), Athlete → NeurIPS (topic and domain change).

From Table 3, we can see that both ALBERT-adjonly and ALBERT-allpairs only experience marginal performance drops over Artist (similar topic), but for Institution and NeurIPS, performance drops substantially, but the relative drop for the adjacency-only method is smaller, suggesting that it captures a more generalised representation of coherence. Indeed, the performance of ALBERT-adjonly in the cross-domain setting is superior or competitive with that for ALBERT-allpairs in the in-domain setting except for PMR over Institution, demonstrating the effectiveness of our training method.

3.6 Evaluation over Multi-document Summarisation

For multi-document summarisation, extractive document summarisation models extract sentences from different documents, not necessarily in an order which maximises discourse coherence. Thus, reordering the extracted sentences is potentially required to maximise the coherence of the extracted text.

We apply our proposed method to multi-document summarisation, in applying ALBERT-allpairs and ALBERT-adjonly to reorder sum-

	λ=0.0	λ=0.3	λ=0.5	λ=0.7	λ=1.0
TextRank	91.28	69.97	55.76	41.55	20.24
allpairs	91.02	70.88	57.45	44.03	23.89
adjonly	**91.94**	**71.76**	**58.30**	**44.85**	**24.67**

Table 4: Coherence scores for reordered summaries. "allpairs" indicates ALBERT-allpairs and "adjonly" indicates ALBERT-adjonly (our model).

maries generated by an extractive multi-document summarisation system. Following Yin et al. (2020), we finetune ALBERT-allpairs and ALBERT-adjonly over 500 reference summaries randomly sampled from a large-scale news summarisation dataset (Fabbri et al., 2019). We then generate extractive summaries from DUC 2004 documents (Task 2) with TextRank (Barrios et al., 2016), and use ALBERT-allpairs and ALBERT-adjonly to reorder the summaries.

To evaluate the coherence of generated summaries, Nayeem and Chali (2017) and Yin et al. (2020) use the weighted sum of cosine similarity and named entity similarity,[3] defined as:

$$\text{Coherence} = \frac{1}{n-1}\sum_{i=1}^{n-1}\text{Sim}(s_i, s_{i+1}),$$

$$\begin{aligned}\text{Sim}(s_i, s_{i+1}) = \lambda * \text{NESim}(s_i, s_{i+1})\\ +(1-\lambda) * \text{Sim}(s_i, s_{i+1}),\end{aligned}$$

where n is the number of sentences, $\text{Sim}(s_i, s_{i+1})$ is the cosine similarity over representations (sum of word embeddings) of adjacent sentences, and $\text{NESim}(s_i, s_{i+1})$ measures the fraction of shared named entities between adjacent sentences. Higher values indicate better performance.

Table 4 shows the results for different λ values (different emphasis on shared named entities). We can see that ALBERT-adjonly achieves higher scores than ALBERT-allpairs and the baseline TextRank for all λ values, once again demonstrating the effectiveness of our method.

4 Conclusion and Future Work

We propose a simple yet effective training method to predict the relative ordering of sentences in a document, based on sentence adjacency and topological sort. Experiments on six datasets from different domains demonstrate the superiority of our proposed method, in addition to results in a cross-domain setting and for multi-document summarisation.

[3] ROUGE score is not used, as it measures content similarity, and does not capture intrinsic text coherence (Koto et al., 2020).

References

Federico Barrios, Federico López, Luis Argerich, and Rosa Wachenchauzer. 2016. Variations of the similarity function of textrank for automated summarization. *arXiv preprint arXiv:1602.03606*.

Regina Barzilay and Mirella Lapata. 2005. Modeling local coherence: An entity-based approach. In *Proceedings of the 43rd Annual Meeting of the Association for Computational Linguistics*, pages 141–148.

Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics*, 34(1):1–34.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3615–3620.

Khyathi Chandu, Eric Nyberg, and Alan W Black. 2019. Storyboarding of recipes: Grounded contextual generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6040–6046.

Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. Neural sentence ordering. *arXiv preprint arXiv:1607.06952*.

Baiyun Cui, Yingming Li, Ming Chen, and Zhongfei Zhang. 2018. Deep attentive sentence ordering network. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4340–4349.

Baiyun Cui, Yingming Li, and Zhongfei Zhang. 2020. BERT-enhanced relational sentence ordering network. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6310–6320.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Micha Elsner, Joseph Austerweil, and Eugene Charniak. 2007. A unified local and global model for discourse coherence. In *Human Language Technologies 2007: The Conference of the North American*

Chapter of the Association for Computational Linguistics; Proceedings of the Main Conference, pages 436–443.

Micha Elsner and Eugene Charniak. 2011. Extending the entity grid with entity-specific features. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 125–129.

Alexander Fabbri, Irene Li, Tianwei She, Suyi Li, and Dragomir Radev. 2019. Multi-news: A large-scale multi-document summarization dataset and abstractive hierarchical model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1074–1084.

Angela Fan, Mike Lewis, and Yann Dauphin. 2019. Strategies for structuring story generation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2650–2660.

Jingjing Gong, Xinchi Chen, Xipeng Qiu, and Xuanjing Huang. 2016. End-to-end neural sentence ordering using pointer network. *arXiv preprint arXiv:1611.04953*.

Camille Guinaudeau and Michael Strube. 2013. Graph-based local coherence modeling. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 93–103.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Junjie Hu, Yu Cheng, Zhe Gan, Jingjing Liu, Jianfeng Gao, and Graham Neubig. 2020. What makes a good story? Designing composite rewards for visual storytelling. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 7969–7976.

Fajri Koto, Jey Han Lau, and Timothy Baldwin. 2020. FFCI: A framework for interpretable automatic evaluation of summarization. *arXiv preprint arXiv:2011.13662*.

Pawan Kumar, Dhanajit Brahma, Harish Karnick, and Piyush Rai. 2020. Deep attentive ranking networks for learning to order sentences. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 8115–8122.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2020. ALBERT: A lite BERT for self-supervised learning of language representations. In *Proceedings of the 8th International Conference on Learning Representations*.

Mirella Lapata. 2006. Automatic evaluation of information ordering: Kendall's tau. *Computational Linguistics*, 32(4):471–484.

Jiwei Li and Eduard Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2039–2048.

Jiwei Li and Dan Jurafsky. 2017. Neural net models of open-domain discourse coherence. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 198–209.

Xia Li, Minping Chen, Jianyun Nie, Zhenxing Liu, Ziheng Feng, and Yingdan Cai. 2018. Coherence-based automated essay scoring using self-attention. In *Proceedings of the 17th China National Conference on Computational Linguistics, CCL 2018, and the 6th International Symposium on Natural Language Processing Based on Naturally Annotated Big Data, NLP-NABD 2018*, pages 386–397.

Lajanugen Logeswaran, Honglak Lee, and Dragomir Radev. 2018. Sentence ordering and coherence modeling using recurrent neural networks. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5285–5292.

Mohsen Mesgar and Michael Strube. 2016. Lexical coherence graph modeling using word embeddings. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1414–1423.

Ramesh Nallapati, Feifei Zhai, and Bowen Zhou. 2017. Summarunner: A recurrent neural network based sequence model for extractive summarization of documents. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*, pages 3075–3081.

Mir Tafseer Nayeem and Yllias Chali. 2017. Extract with order for coherent multi-document summarization. In *Proceedings of TextGraphs-11: the Workshop on Graph-based Methods for Natural Language Processing*, pages 51–56.

Byungkook Oh, Seungmin Seo, Cheolheon Shin, Eunju Jo, and Kyong-Ho Lee. 2019. Topic-guided coherence modeling for sentence ordering by preserving global and local information. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2273–2283.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, pages 1532–1543.

Shrimai Prabhumoye, Ruslan Salakhutdinov, and Alan W Black. 2020. Topological sort for sentence ordering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2783–2792.

Robert Endre Tarjan. 1976. Edge-disjoint spanning trees and depth-first search. *Acta Informatica*, 6(2):171–185.

Yi Tay, Minh C. Phan, Luu Anh Tuan, and Siu Cheung Hui. 2018. SkipFlow: Incorporating neural coherence features for end-to-end automatic text scoring. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*, pages 5948–5955.

Tianming Wang and Xiaojun Wan. 2019. Hierarchical attention networks for sentence ordering. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence (AAAI-19)*, volume 33, pages 7184–7191.

Dejing Xu, Jun Xiao, Zhou Zhao, Jian Shao, Di Xie, and Yueting Zhuang. 2019a. Self-supervised spatiotemporal learning via video clip order prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10334–10343.

Peng Xu, Hamidreza Saghir, Jin Sung Kang, Teng Long, Avishek Joey Bose, Yanshuai Cao, and Jackie Chi Kit Cheung. 2019b. A cross-domain transferable neural coherence model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 678–687.

Yongjing Yin, Fandong Meng, Jinsong Su, Yubin Ge, Linfeng Song, Jie Zhou, and Jiebo Luo. 2020. Enhancing pointer network for sentence ordering with pairwise ordering predictions. In *Proceedings of the Thirty-Fourth AAAI Conference on Artificial Intelligence (AAAI-20)*, pages 9482–9489.

Yongjing Yin, Linfeng Song, Jinsong Su, Jiali Zeng, Chulun Zhou, and Jiebo Luo. 2019. Graph-based neural sentence ordering. In *Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence, IJCAI*, pages 5387–5393.